Can a Random Forest Model predict the presence of Heart Disease?

By Jan Möhle

Outline

1. Data set description
2. Variables
3. Methodology
4. Results
5. Comparison with two other classification models
6. Conclusion

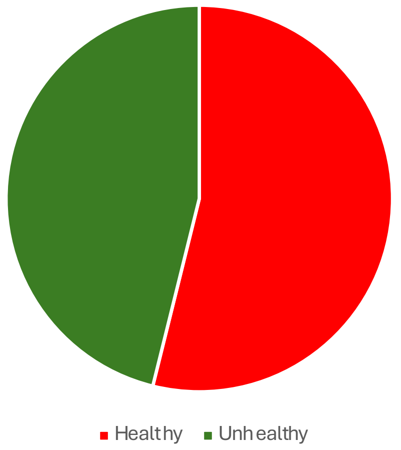
Data set description

- Data set from Cleveland Clinic in Cleveland, Ohio from 1989
  - Contains clinical data about patients with and without heart disease (coronary artery disease)
  - Often used by ML researchers
- First used in: “International application of a new probability algorithm for the diagnosis of coronary artery disease” by Robert Detrano et al. in 1989
  - Goal: testing probability algorithms to predict heart disease
  - Findings: prediction works in general good, but several algorithms overpredicted the probability of heart disease

Outcome variable

Variable Name	Description	Variable Type	Values
hd	diagnosis of heart disease	binary	0: no 1: yes

- 297 observations
- 13 features (explanatory variables)
- Balanced sample:
    - Healthy: 160 (54%)
    - Unhealthy: 137 (46 %)



Explanatory variables

Variable Name	Description	Variable Type	Values
age	age in years	continues values	[29, 77]
sex	gender	binary	0: female 1: male
cp	chest pain type	ordered factor	1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
trestbps	resting blood pressure (in mm Hg on admission to the hospital)	continues values	[94, 200]
chol	serum cholestoral in mg/dl	continues values	[126, 564]
fbs	fasting blood sugar > 120 mg/dl	binary	0: no 1: yes
thalach	maximum heart rate achieved	continues values	[71, 202]
exang	exercise induced angina	binary	0: no 1: yes

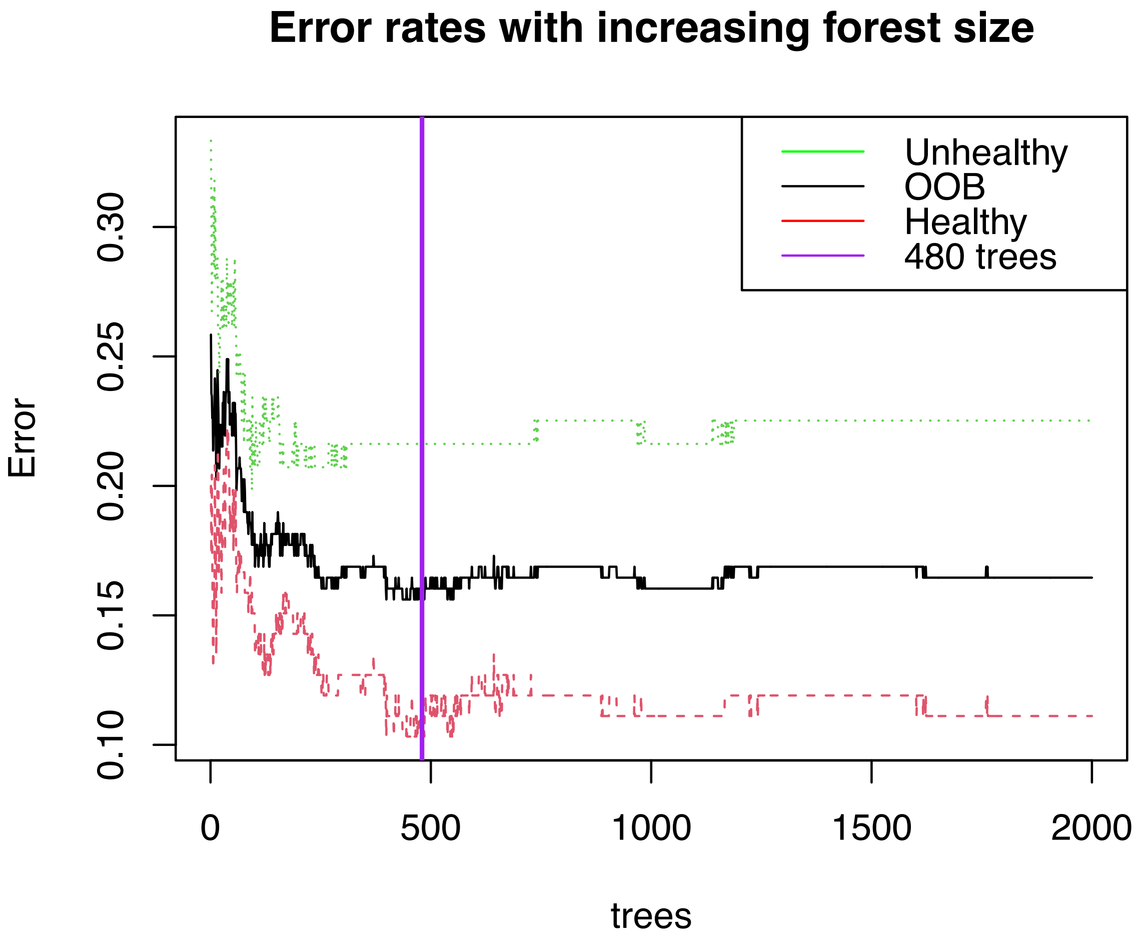
Explanatory variables

Variable Name	Description	Variable Type	Values
restecg	resting electrocardiographic results	ordered factor	1: normal 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 3: showing probable or definite left ventricular hypertrophy by Estes' criteria
oldpeak	ST depression induced by exercise relative to rest	continues values	[0.0, 6.2]
slope	the slope of the peak exercise ST segment	ordered factor	1: upsloping 2: flat 3: downsloping
ca	number of major vessels colored by flourosopy	ordered factor	{0, 1, 2, 3}
thal	thalium heart scan	ordered factor	1: normal 2: reversable 3: fixed defect

Methodology

- Using Random Forest as classifier
- Set seed for every process that includes randomness
    - ➢Reproducibility
- Cross-Validation for comparison with other methods
    - ➢Train dataset: 80% of observations (237)
    - ➢Test dataset: 20% of observations (60)
- Choice of number of trees based on error plot
- Choice of number of features to consider in each tree based on optimal OOB error

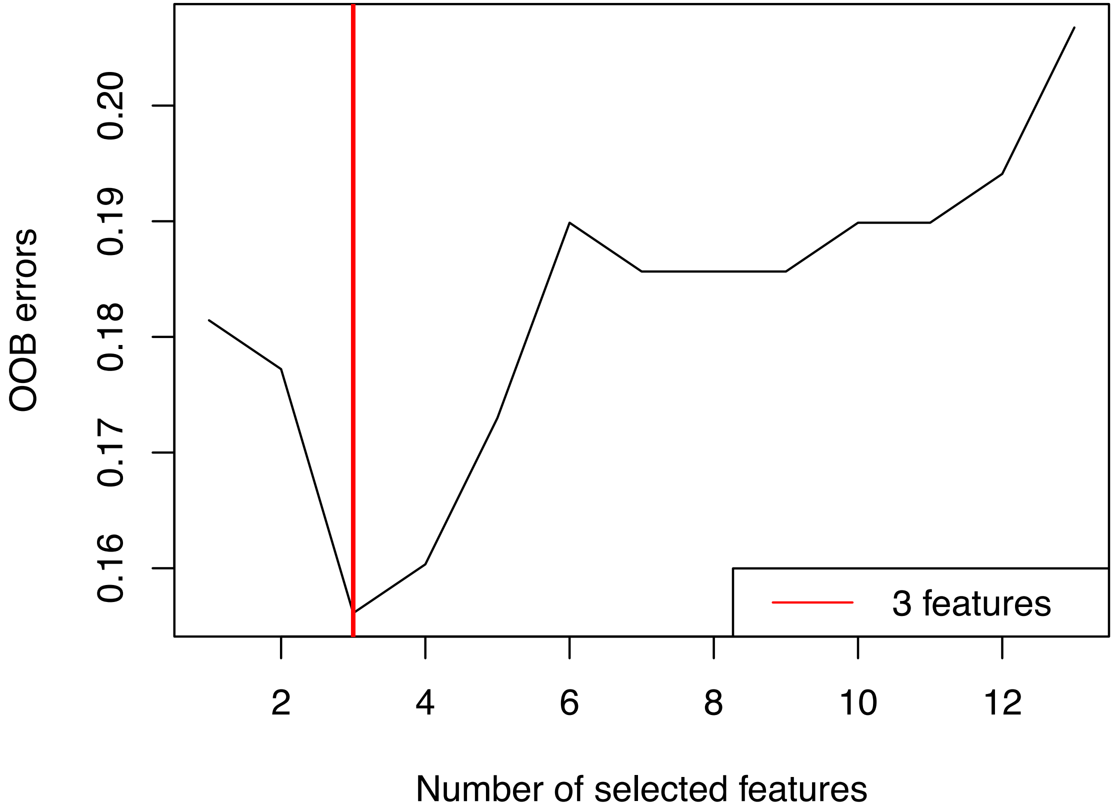
Choice of number of trees



- All error rates seem to stabilize with around 400 trees
- Default setting in R: 500

➢480 trees seems to be a good choice (smallest error)

Choice of number of features

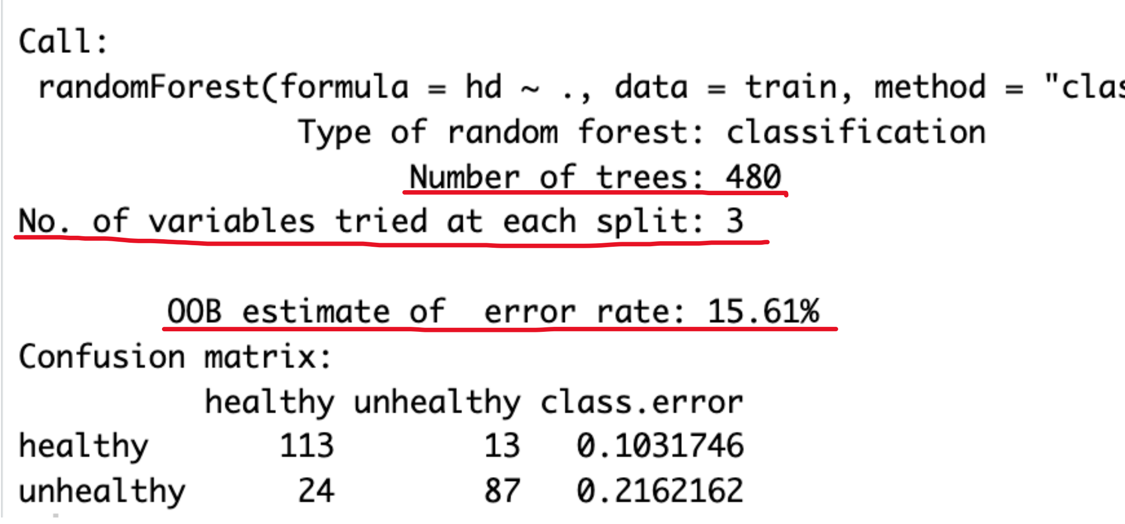


- Trying 1 to 13 with for loop

➢ choosing option with lowest OOB error

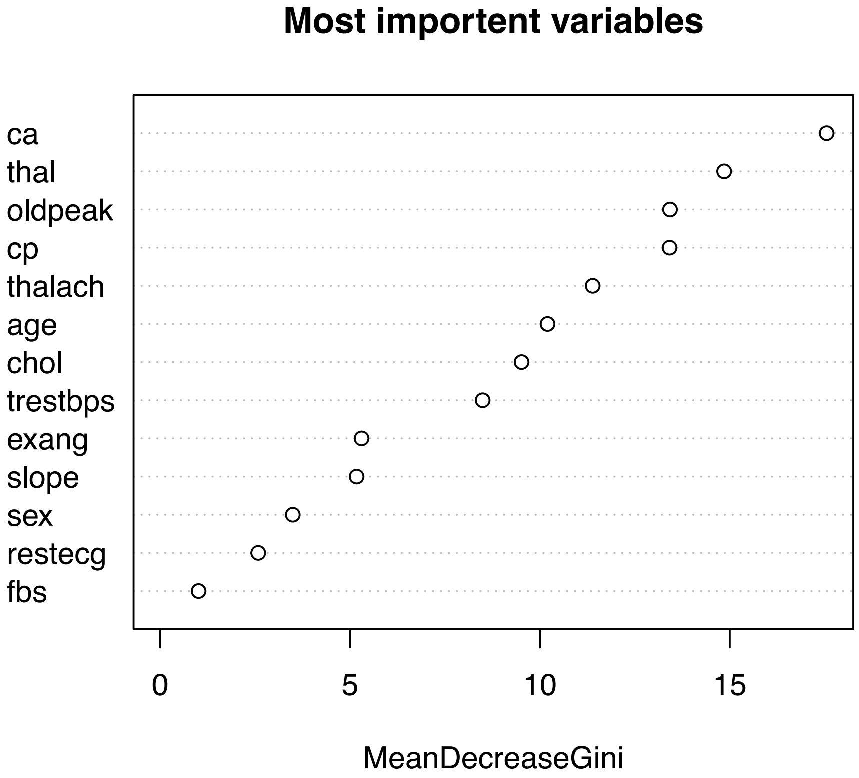
- Choosing 3 values for model
  - Minimal OOB error
  - Close to rule of thumb: $\sqrt{features} = 3.6$

Final forest and results



```
Call:
 randomForest(formula = hd ~ ., data = train, method = "class",      ntree = 480, mtry = 3)
               Type of random forest: classification
                     Number of trees: 480
No. of variables tried at each split: 3

        OOB estimate of  error rate: 15.61%
Confusion matrix:
          healthy unhealthy class.error
healthy       113        13   0.1031746
unhealthy      24        87   0.2162162
```



In train data:

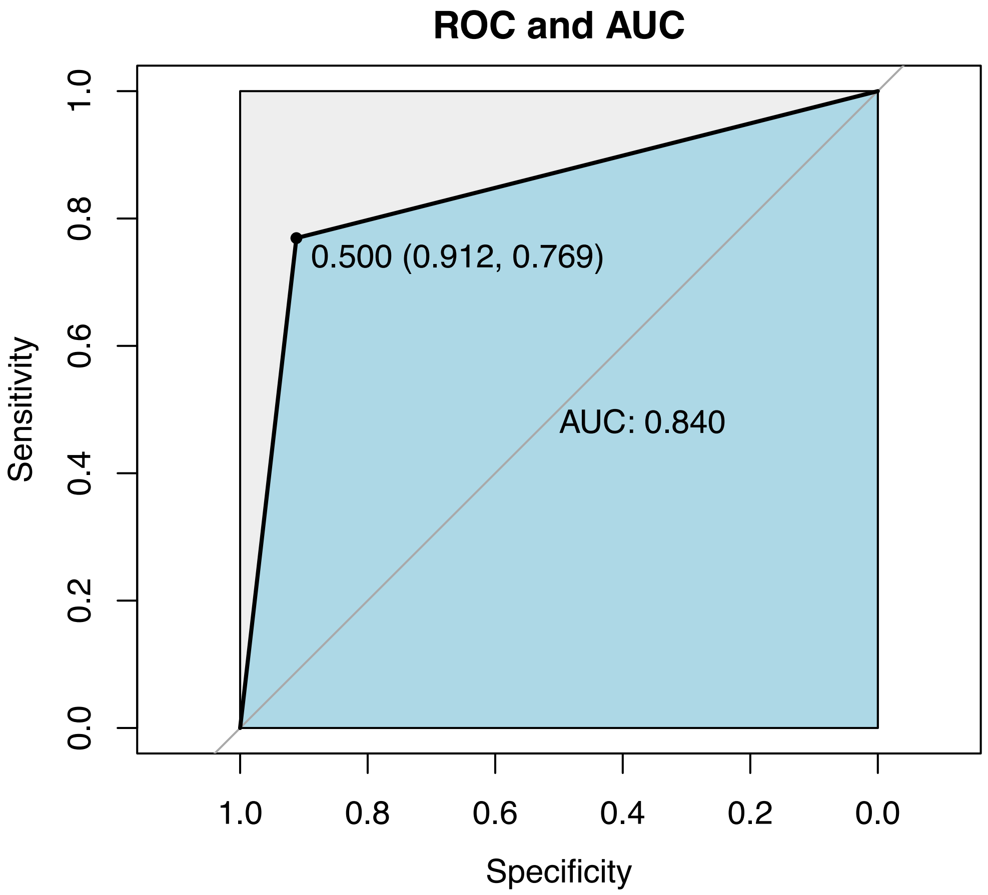
- $RP \approx 85\%$
- $TNR \approx 90\%$
- $TPR \approx 78\%$

Predictions with test data

Test data	**Predicted**	
Observed	Healthy	Unhealthy
Healthy	31	3
Unhealthy	6	20

In test data:

- $RPR \approx 85\%$
- $TNR \approx 91\%$
- $TPR \approx 77\%$



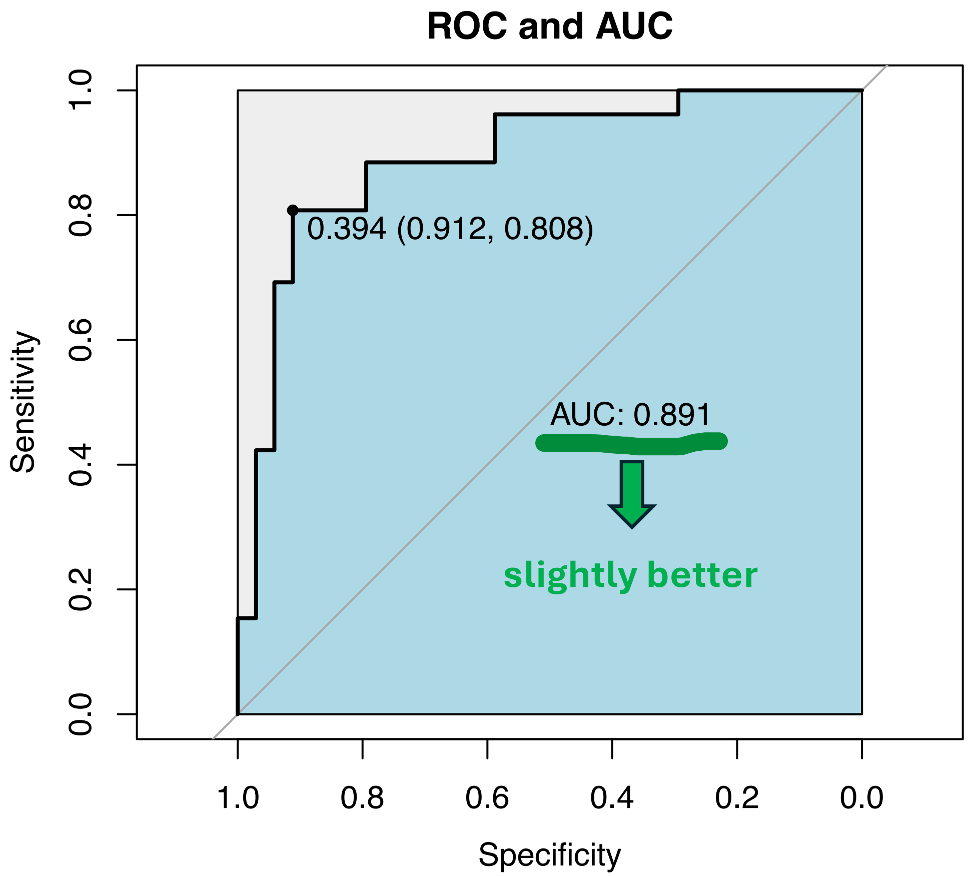
Comparing results with two other classification models

Logistic Regression prediction results

Test data	Predicted	
Observed	Healthy	Unhealthy
Healthy	31	3
Unhealthy	5	21

In test data:

- $RPR \approx 87\%$ → slightly better
- $TNR \approx 91\%$
- $TPR \approx 81\%$ → slightly better

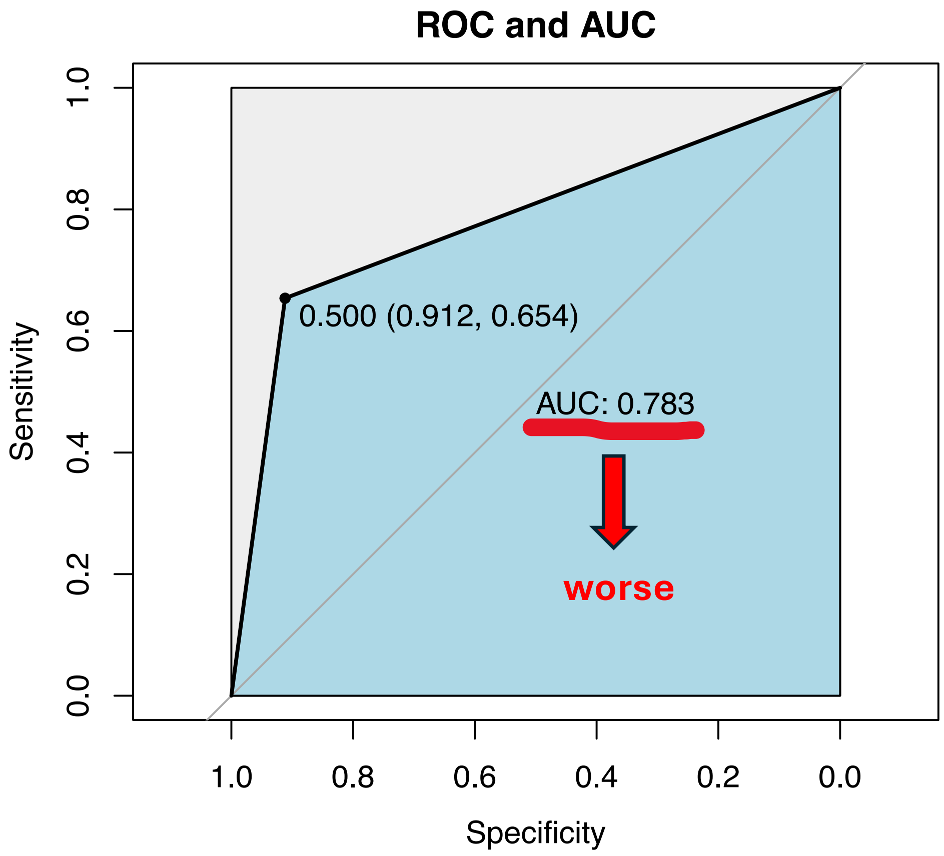


Decision Tree prediction results

Test data	Predicted	
Observed	Healthy	Unhealthy
Healthy	31	3
Unhealthy	9	17

In test data:

- $RPR \approx 80\%$ → worse
- $TNR \approx 91\%$
- $TPR \approx 65\%$ → worse



Conclusion

- Prediction of heart disease based on medical measurement with random forest possible
- Logistic Regression slightly better in that case (higher TPR)
  - ➢Question: Logistic Regression in general better for that problem?
    - ➢Only slight difference, so probably not
- Decision tree worse in that case (lower TPR)

❖Restrictions of analysis:

- Small data set
  - Random Forest tends to work better with more observations

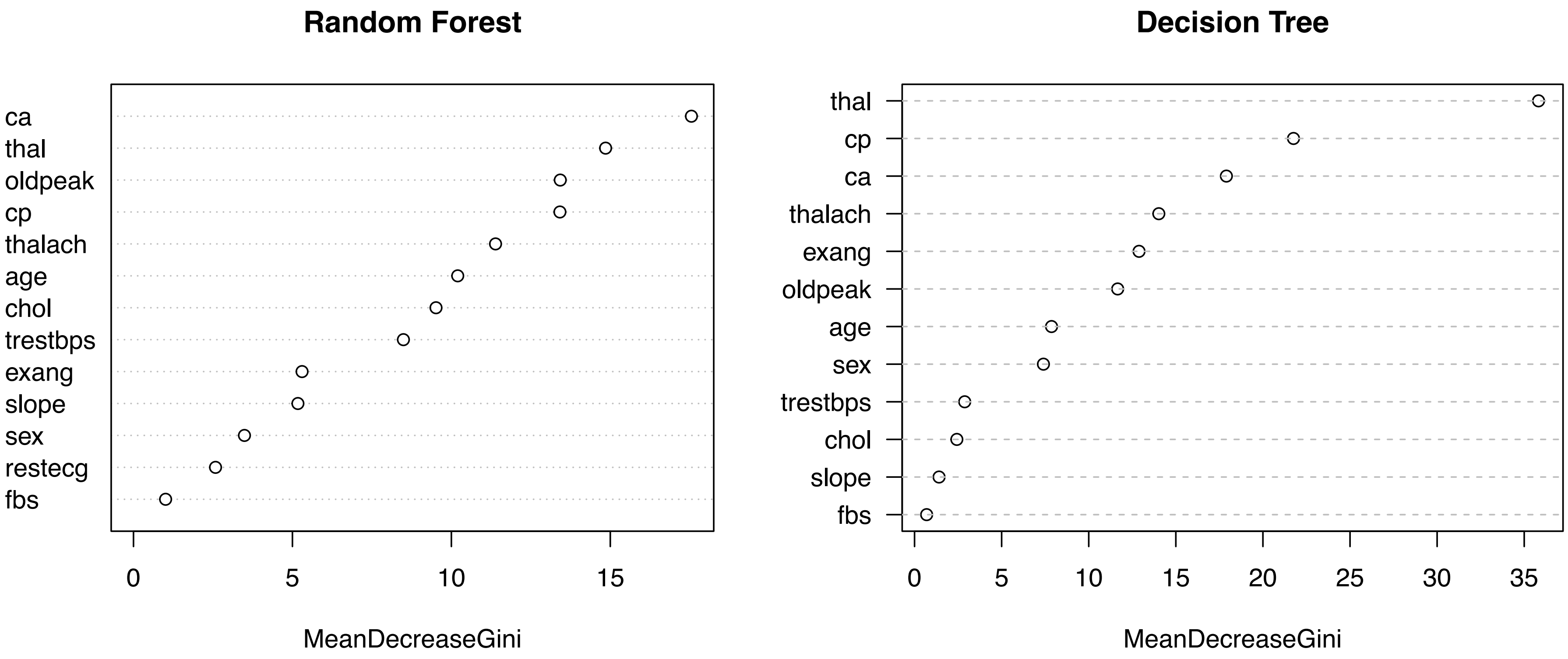
Sources

- Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, Guppy KH, Lee S, Froelicher V. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol. 1989 Aug 1;64(5):304-10. doi: 10.1016/0002-9149(89)90524-9. PMID: 2756873.
- Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.
- Direct link to data set: http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

Thank you for listening!

Appendix

Comparing variable importance: Tree and Forest


Random Forest
ca
thal
oldpeak
cp
thalach
age
chol
trestbps
exang
slope
sex
restecg
fbs
MeanDecreaseGini
Decision Tree
thal
cp
ca
thalach
exang
oldpeak
age
sex
trestbps
chol
slope
fbs
MeanDecreaseGini